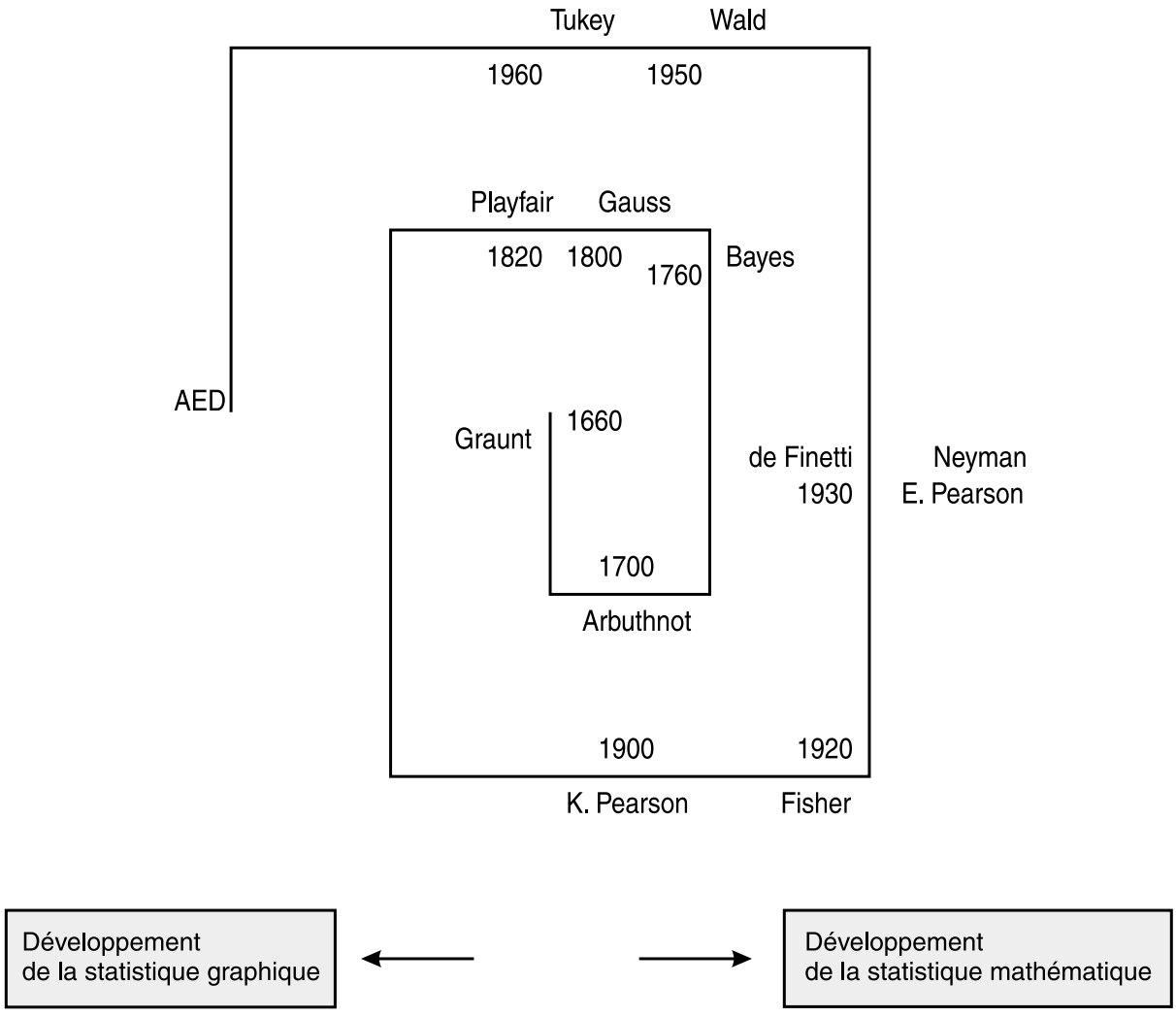
Le chemin de la statistique, d'après Peter J. HUBER



AED mode d'emploi

L'objectif de l'analyse exploratoire des données est de découvrir des structures, des groupes, des relations..., en se gardant autant que faire se peut des présupposés. Cette approche s'appuie sur un équipement adéquat, en l'occurrence un logiciel statistique perfectionné, intégrant des fonctions de représentation graphique adaptées. À tous les stades de l'analyse, l'explorateur devra en effet pouvoir visualiser ses données et l'effet des traitements qu'il leur aura appliqués, cela en mode interactif et sous différents points de vue. Réexpression des données, résistance et robustesse des indicateurs et des procédures et analyse fine des « restes » (en analyse exploratoire, on préfère parler de restes plutôt que de résidus) sont les trois autres fondamentaux de l'AED.

Mettre en scène les données

L'explorateur, pour se faire une idée des structures que recèlent ses données, doit pouvoir s'appuyer sur des représentations graphiques expressives, les plus complètes possibles, matérialisant la façon dont se distribuent les variables étudiées : *Ideas come from previous exploration more than from lightning strokes*.

À cette fin, Tukey a perfectionné des représentations graphiques déjà existantes, imaginant par exemple le *Stem and Leaf,* cousin de l'histogramme. Il en a également inventé de nouvelles, notamment le *Box Plot* (abréviation de *Box and Whiskers Plot,* qui peut se traduire par boîte à moustaches ou encore boîte à pattes). Parmi les autres représentations graphiques les plus couramment utilisées en analyse exploratoire des données, on citera en particulier les diagrammes de dispersion et matrices de diagrammes de dispersion (pour l'étude des liaisons pouvant exister entre les variables), les *QQplots* et *PPplots* (qui permettent de comparer des distributions observées à des distributions théoriques), les diagrammes de rotation en 3D (pour l'interprétation des plans factoriels).

Graphiques et statistique

Avec trois nombres vous faites une phrase, avec quatre à vingt nombres vous faites un tableau, au delà de vingt nombres vous faites un graphique (Tufte).

Dès l'origine (Quételet, Galton) et pendant très longtemps, la représentation graphique a joué dans l'analyse des données réelles un rôle essentiel. Cette pratique s'est toutefois diluée aux premiers temps de l'informatique (1950-1975), car peu adaptée aux sorties ligne à ligne des imprimantes de l'époque. Pendant cette période, on s'est donc essentiellement appliqué à produire du chiffre. Mais avec les nouveaux outils informatiques et sous l'impulsion de Tukey, la statistique graphique a aujourd'hui retrouvé toute sa respectabilité.

La représentation graphique n'est pas seulement un moyen de regarder les données. Couplée avec l'interactivité, elle devient en effet un puissant outil d'analyse. Par exemple, l'utilisation de marqueurs différents va permettre de discriminer des groupes de données sur un diagramme de dispersion, ajoutant ainsi une troisième dimension à la relation spatiale entre deux variables. On pourra même y faire apparaître une quatrième variable au moyen de la couleur.

Des exemples concrets et commentés sont donnés dans l'article « Visualisation interactive et réexpression des données avec Sas/Insight ».

Réexprimer les données

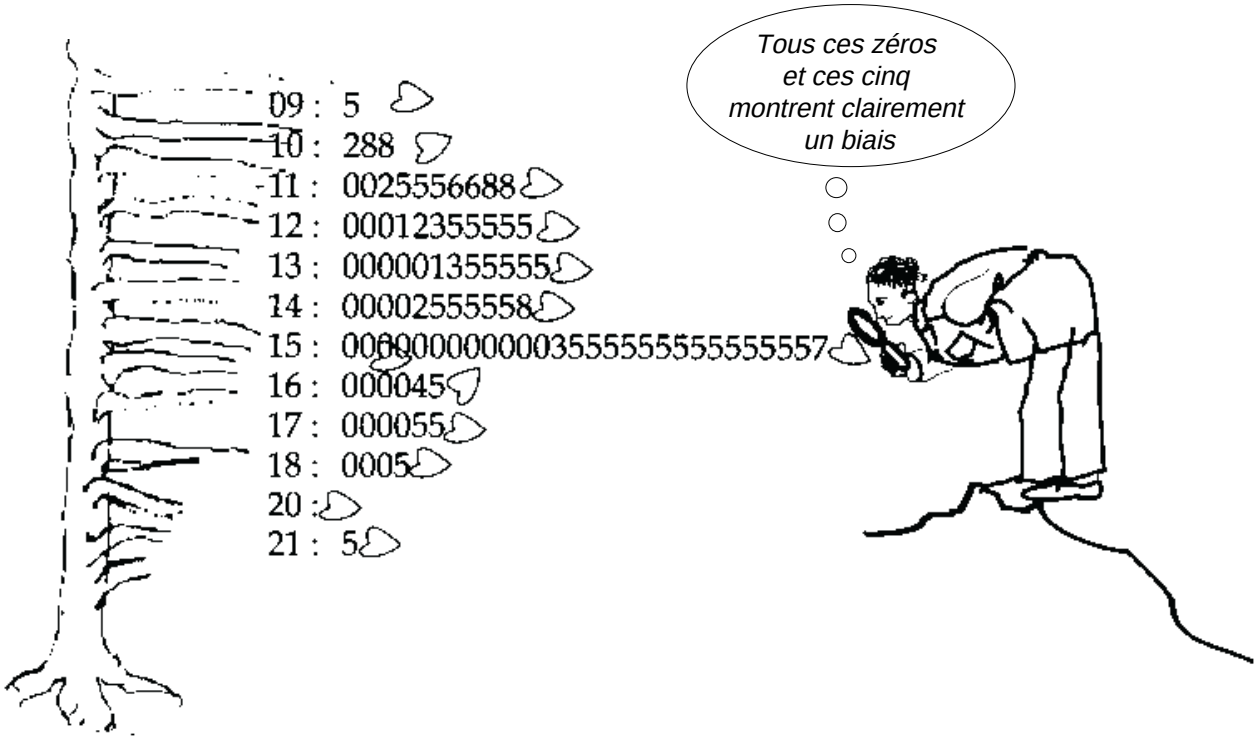
Le novice a souvent tendance à rester attaché à la forme première des données. Or celles-ci ne sont pas « données » (*data are not given*). On peut les adapter, afin d'y découvrir des structures cachées. Mais en la matière, il faut

	Puissance	Transformée	
	etc.	etc.	*Monter l'échelle*
	4	Y^4	
	3	Y^3	
	2	Y^2	
Variable brute	**1**	**Y**	
	1/2	$Y^{1/2}$	
	0	lnY	
	-1/2	$-1/\sqrt{Y}$	
	-1	-1/Y	
	-2	$-1/ Y^2$	
	-3	$-1/ Y^3$	
	etc.	etc.	*Descendre l'échelle*

Échelle de puissance de Tukey
(selon la représentation originale de Horber E.)

Le Stem and Leaf, cousin de l'histogramme

John Tukey a inventé la représentation *Stem and Leaf* (tige et feuille), qui permet de résumer les observations tout en conservant les données individuelles.



Lecture :
sur la 1re ligne, lire la valeur 95 ;
sur la 2e ligne, lire les 3 valeurs 102, 108 et 108 ;
sur la 3e ligne, lire les 10 valeurs 110, 110, 112, etc.

Réalisation d'après une idée de Larry Gonick & Woollcott Smith
(« The carton guide of statistics », Harper Perennial, 1993)

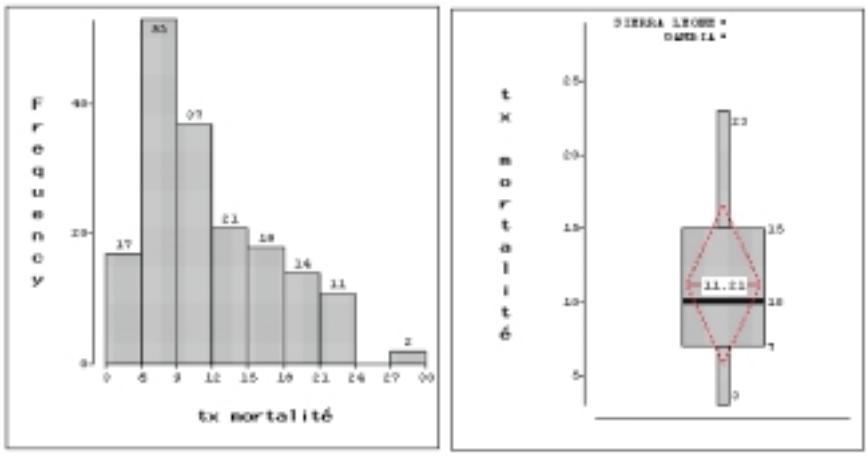
Les Box Plots, une mine d'informations

La variable ici étudiée est le taux de mortalité (pour mille habitants), dont on a relevé la valeur dans 173 pays (53 en Afrique, 39 en Amérique, 44 en Asie, 28 en Europe et 9 en Océanie).

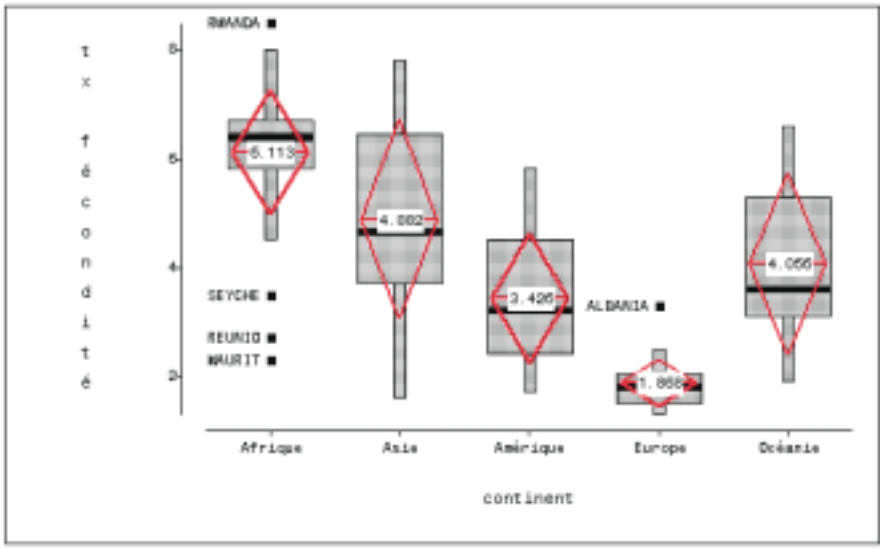
À gauche est donné un histogramme classique. À droite un Box Plot, qui apporte de nombreuses informations supplémentaires :

- la hauteur de la boîte (rectangle central) détermine l'étendue de la partie centrale de la distribution (taux de mortalité compris entre les quartiles Q1 et Q3, valeurs 7 et 15) ;
- la bande horizontale matérialisée à l'intérieur de cette même boîte indique la position de la médiane (valeur 10) ;
- l'étendue des queues de distribution hors données atypiques est déterminée par la hauteur des moustaches, étant précisé que l'extrémité de la moustache inférieure est ici fixée par la plus petite valeur (= 3) supérieure ou égale à Q1-1,5 (Q3 - Q1), celle de la moustache supérieure par la plus grande valeur (= 23) inférieure ou égale à Q3 +1,5 (Q3 - Q1) ;
- de part et d'autre des moustaches sont mises en évidence les observations atypiques, avec valeur inférieure à Q1 - 1,5 (Q3 - Q1) ou supérieure à Q3 + 1,5 (Q3 - Q1), ici la Gambie et la Sierra Leone, où le taux de mortalité est particulièrement élevé.

On peut également faire apparaître les valeurs de la moyenne et de l'écart-type. Dans la représentation ci-dessous, la petite diagonale du losange superposé à la boîte indique la position de la moyenne (valeur 11,21), sa grande diagonale, de longueur 2 σ, permet d'apprécier la valeur de l'écart-type.



Intéressons-nous à présent à la distribution du taux de fécondité (nombre d'enfants par femme). Par simple juxtaposition des Box Plots relatifs aux différents continents, on voit tout de suite que c'est en Europe que la concentration est de très loin la plus forte (autour d'une moyenne faible et très proche de la médiane), qu'il existe dans ce même continent un pays atypique (l'Albanie), que c'est en Asie que la distribution est la plus dispersée, et cetera, et cetera.
Ce type de représentation constitue une excellente introduction visuelle à l'analyse de la variance.



aussi savoir raison garder... Aussi Tukey a-t-il proposé à l'usage de l'explorateur une échelle de transformations raisonnable et raisonnée.

Dans la pratique, deux utilisations très répandues de la transformation (ou réexpression, selon le terme de Tukey) des données sont la symétrisation de la distribution d'une variable et la linéarisation d'une liaison entre deux variables.

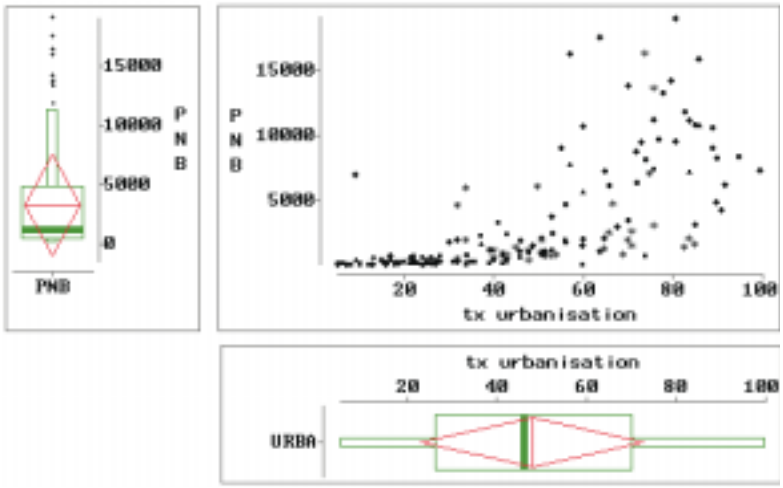
Si par exemple la courbe représentative de la distribution d'une variable est assymétrique avec une queue plus étalée à droite, peut-être pourra-t-on découvrir, en s'intéressant à son logarithme (puissance 0 dans l'échelle de Tukey), que cette variable suit la loi log-normale (un cas bien connu est celui de la distribution du *revenu*). De même, si l'on cherche à établir la liaison, a priori non linéaire, pouvant exister entre deux variables, on pourra utilement procéder à des essais de transformation successifs sur l'une ou l'autre de ces deux variables, en montant ou descendant l'échelle de Tukey. Peut-être mettra-t-on alors au jour une liaison linéaire entre la première variable et une transformée de la seconde.

Avec les nouveaux outils de la micro, réaliser ces transformations est devenu très facile et « compréhensible ».

De l'utilité de la réexpression des données

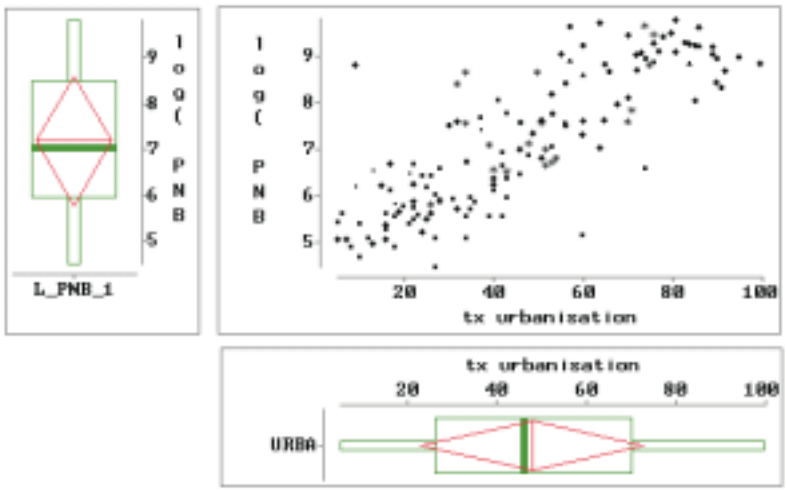
Les variables ici étudiées sont le PNB et le taux d'urbanisation, le nombre d'observations (nombre de pays observés) est égal à 173.

Dans cette première fenêtre graphique sont regroupés deux Box Plots, le premier relatif à la distribution du PNB, le second à la distribution du taux d'urbanisation, ainsi qu'un diagramme de dispersion présentant la façon dont se répartissent les 173 pays observés suivant ces deux mêmes variables.



La même configuration est reprise ci-dessous, à cet important détail près : la variable PNB a été transformée en logPNB (puissance zéro sur l'échelle de Tukey).

On constate alors, d'une part que la distribution de logPNB est quasi symétrique, d'autre part qu'il semble y avoir une liaison linéaire entre cette variable et le taux d'urbanisation.



Résistance et robustesse

Dans le langage courant, résistance et robustesse désignent des qualités très voisines. S'agissant de la chose statistique, on admettra, comme Velleman, que le qualificatif de résistant doit plutôt s'appliquer à un indicateur, celui de robuste à une méthode.

Un indicateur sera dit résistant s'il est peu sensible à l'effet des données atypiques : ainsi la médiane est un indicateur de tendance centrale résistant, mais la moyenne ne l'est pas.

Une méthode sera dite robuste si elle est peu sensible à des déviations du corpus d'hypothèses qui en principe la sous-tend : ainsi, la pourtant si précieuse régression linéaire basée sur les moindres carrés n'est pas une méthode robuste.

Dispersion d'une distribution

Un principe général, couramment admis, est que « les données sont toujours normales en leur milieu ». Encore faut-il déterminer où se situe le milieu... En phase exploratoire, on évitera donc de se focaliser sur la moyenne, très sensible à l'effet des données atypiques. On préfèrera au contraire s'intéresser à des indicateurs de tendance centrale résistants, tels que la médiane, les moyennes tronquées ou winsorisées, la *midmean* ou encore la *trimean.*

Une autre règle d'or de l'AED est qu'une distribution doit être étudiée sur toute sa longueur. Confronté par exemple à une courbe en forme de cloche, l'explorateur ne se laissera pas obnubiler par l'hypothèse de normalité, souvent illusoire, et privilégiera la mise en œuvre de méthodes non paramétriques. Il se préoccupera bien sûr de la forme plus ou moins aplatie (Kurtosis) du milieu de la courbe, mais accordera une attention prioritaire aux aspects de symétrie. Les queues de distribution notamment doivent faire l'objet d'un examen approfondi, et c'est bien dans cette perspective que Tukey avait imaginé le *Box Plot.*

Ajustement d'une liaison linéaire

Robustesse toujours, Tukey a inventé, en matière d'ajustement linéaire, un substitut à la droite des moindres carrés, la *Tukey's Line*, à construire, dans un graphique cartésien (X, Y), selon le schéma ci-dessous :

- on découpe le nuage de points selon la valeur de X en trois groupes de taille identique (même nombre de points dans chacun) ;
- on calcule les valeurs médianes de X et de Y pour les premier et troisième groupes ; les valeurs ainsi calculées déterminent deux points (le point médian du premier groupe et celui du troisième) ;
- on trace la droite passant par ces deux points (droite des médianes croisées) ;
- la *Tukey's Line* est la parallèle à cette droite qui partage le nuage de points en deux groupes de taille identique.

Analyser les restes

La démarche exploratoire est par nature itérative. Prenons l'exemple de la distribution des taux de fécondité en Afrique, dont le *Box Plot* est présenté dans l'encadré de la page 19. On remarque instantanément que quatre pays se distinguent : le Rwanda, les Seychelles, la Réunion et l'Île Maurice. L'explorateur va essayer d'expliquer ces particularités en échafaudant des hypothèses : les trois pays aux taux de fécondité les plus bas sont des îles de l'océan Indien ; ce point commun est-il un élément d'explication ? Si oui, pourquoi Madagascar n'est-il pas atypique ? etc. Lorsqu'il aura trouvé une explication plausible à cette particularité, l'explorateur va s'intéresser aux autres pays, au « reste » de la distribution, qu'il représentera par un autre graphique, graphique qui révèlera peut être d'autres particularités qu'il expliquera, etc.

L'idée de base de l'exploration est en effet de décomposer les données selon une structure, les points atypiques dans notre exemple, et un reste, dont on recherchera à nouveau la décomposition en structure + reste, et ainsi de suite.

Pour transcrire cet enchaînement, on trouve dans la littérature la séquence d'« équations » suivante :

Data = Smooth_1 + Rough_1
Rough_1 = Smooth_2 + Rough_2
Etc.

Les données sont donc divisées en une structure souvent simple, interprétable, et qualifiée de « lisse » dans la mesure où elle ne pose plus de problème majeur, et en un reste, non encore étudié et interprété, qualifié de « rugueux ».

Remarquons que le reste n'a aucune raison d'être « petit ». Ainsi, dans notre exemple, la structure exhibée concerne quatre pays, et le reste, qui concerne 49 pays, contient sans doute la majeure partie de l'information. C'est pourquoi les explorateurs ne parlent pas de « résidu », ce terme faisant souvent référence, en statistique classique, à quelque chose de petit.

Cette stratégie itérative d'analyse est toujours utilisée en exploration, où les restes font l'objet d'une attention toute particulière.

Sophie DESTANDAU,
Dominique LADIRAY
et **Monique LE GUEN**

Les moyennes tronquées (*Trimmed Means*) sont des moyennes arithmétiques calculées après élimination des valeurs extrêmes (la plus petite et la plus grande, les 2 plus petites et les 2 plus grandes, etc.).

Les moyennes winsorisées (*Winsorized Means*) ont été introduites par Winsor dans les années 1940. Au lieu de supprimer les valeurs extrêmes, Winsor les remplace par la valeur la plus proche non suspecte.

La midmean est la moyenne des valeurs supérieures ou égales au premier quartile (Q1) et inférieures ou égales au troisième (Q3).

La **trimean** est égale à (Q1+2Q2+Q3) / 4, où Q2 désigne la médiane.

L’analyse de données à la française dans la typologie de Tukey

Sous la dénomination générique d’analyse de données à la française sont regroupées l’ACP (analyse en composantes principales), l’AFC (analyse factorielle des correspondances) et l’ACM (analyse des correspondances multiples) ainsi que les méthodes de classification, soit autant d’approches relevant de l’analyse exploratoire multidimensionnelle au sens de Tukey. C’est d’ailleurs bien ainsi que les statisticiens français les présentent aujourd’hui.

Cf. Lebart L., Morineau A., Piron M. : Statistique exploratoire multidimensionnelle, 1997.